Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Adv. Methods

Module 4 – Advanced Analytics - Theory and Methods

Module 4: Advanced Analytics – Theory and Methods

Part 8: Text Analysis

During this lesson the following topics are covered:

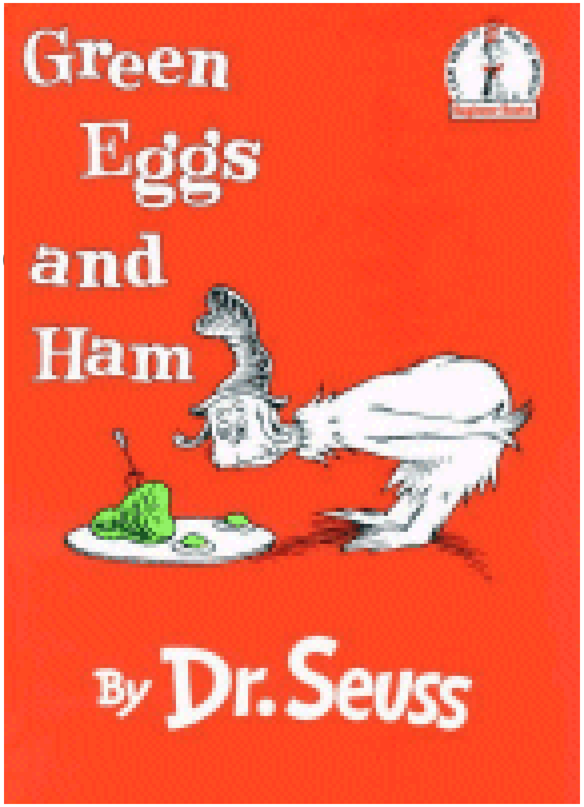
- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
  - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Use of regular expressions in parsing text
- Metrics used to measure the quality of search results
  - Relevance with tf-idf, precision and recall

Text Analysis

The processing and representation of text for analysis and learning tasks

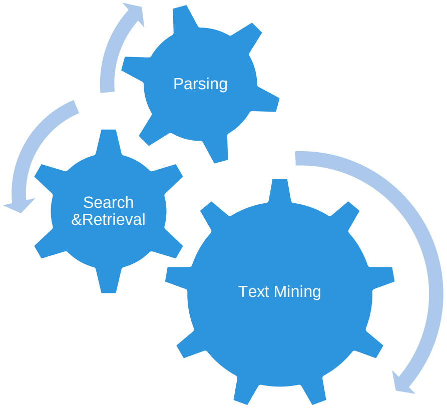
The main challenges in text analysis

- **High-dimensionality**
  - Every distinct term is a dimension
  - When analyzing a document every possible w represents a dimension.
  - *Green Eggs and Ham*: A 50-D problem!
- **Data is Un-structured**

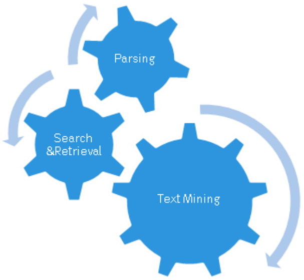


Text Analysis – Problem-solving Tasks

- Parsing
  - Impose a structure on the unstructured/semi-structured text for downstream analysis
- Search/Retrieval
  - Which documents have this word or phrase?
  - Which documents are about this topic or this entity?
- Text-mining
  - "Understand" the content
  - Clustering, classification
- Tasks are not an ordered list
  - Does not represent process
  - Set of tasks used appropriately depending on the problem addressed
- Usually you start with parsing then do either search or text mining

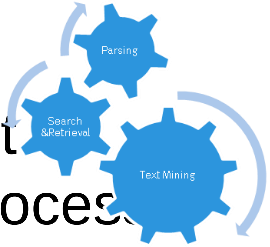


Example: Brand Management



- Acme currently makes two products
  - bPhone
  - bEbook
- They have lots of competition. They want to maintain their reputation for excellent products and keep their sales high.
- What is the buzz on Acme?
  - Search for mentions of Acme products
    - Twitter, Facebook, Review Sites, etc.
  - What do people say?
    - Positive or negative?
    - What do people think is good or bad about the products?

Buzz Tracking: The Process



The tasks carried out for the tracking vs. the corresponding text analysis tasks associated with the established buzz tracking process

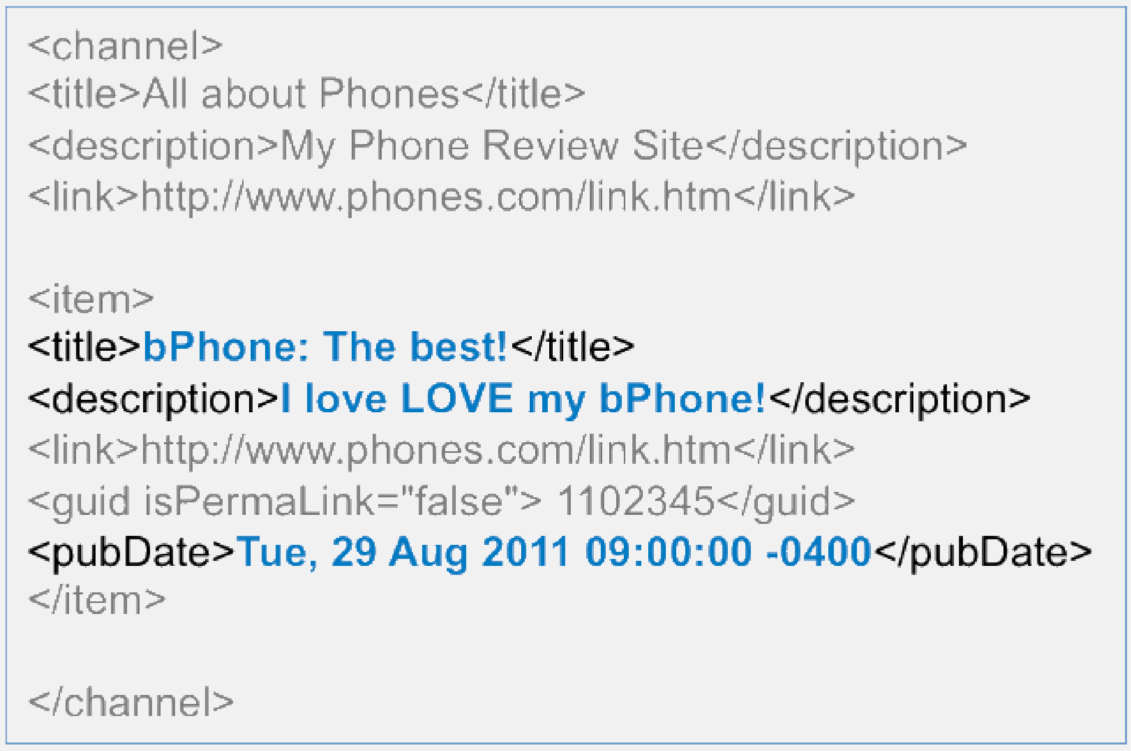
1. Monitor social networks, review sites for mentions of our products.	**Parse** the data feeds to get actual content. Find and filter the raw text for product names (Use **Regular Expression**).
2. Collect the reviews.	Extract the relevant raw text. Convert the raw text into a suitable **document representation**. **Index** into our review **corpus**.
3. Sort the reviews by product.	**Classification** (or **"Topic Tagging"**)
4. Are they good reviews or bad reviews? We can keep a simple count here, for trend analysis.	**Classification** (sentiment analysis)
5. Marketing calls up and reads selected reviews in full, for greater insight.	**Search/Information Retrieval**.

Parsing the Feeds

Parsin
g

1. Monitor social networks, review sites for mentions of our products

- Impose structure on semi-structured data.
- We need to know where to look for what we are looking for.



```
<channel>
<title>All about Phones</title>
<description>My Phone Review Site</description>
<link>http://www.phones.com/link.htm</link>

<item>
<title>bPhone: The best!</title>
<description>I love LOVE my bPhone!</description>
<link>http://www.phones.com/link.htm</link>
<guid isPermaLink="false"> 1102345</guid>
<pubDate>Tue, 29 Aug 2011 09:00:00 -0400</pubDate>
</item>

</channel>
```

Regular Expressions

1. Monitor social networks, review sites for mentions of our products

- Regular Expressions (regexp) are a means for finding words, strings or particular patterns in text.
- A match is a Boolean response. The basic use is to ask "does this regexp match this string?"

regexp	matches	Note
b[P|p]hone	bPhone, bphone	Pipe "|" means "or"
bEbo*k	bEbk, bEbok, bEbook, bEboook …	"*" matches 0 or more repetitions of the preceding letter
^I love	A line starting with "I love"	"^" means start of a string
Acme$	A line ending with "Acme"	"$" means the end of a string

Extract and Represent Text

2. Collect the reviews

Document Representation:

A structure for analysis

- **"Bag of words"**
  - common representation
  - A vector with one dimension for every unique term in space
    - **term-frequency (tf)**: number times a term occurs
  - Good for basic search, classification
- Reduce Dimensionality
  - Term Space – not ALL terms
    - no stop words: "the", "a"
    - often no pronouns
  - Stemming
    - "phone" = "phones"

"I love LOVE my bPhone!"

Convert this to a vector in the term space:

acme	0
bebook	0
bPhone	1
fantastic	0
love	2
slow	0
terrible	0
terrific	0

Parsin
g

Document Representation - Other Features

2. Collect the reviews

- Feature:
  - Anything about the document that is used for search or analysis.
- Title
- Keywords or tags
- Date information
- Source information
- Named entities

Representing a Corpus (Collection of Documents)

Parsin
g

2. Collect the reviews

- It is important that we not only create a representation of the document but we also need to represent a corpus.
- Reverse index
  - For every possible feature, a list of all the documents that contain that feature
  - “Reverse index” provides a way of keeping track of list of all documents that contain a specific feature and for every possible feature.
- Corpus metrics
  - Volume
  - Corpus-wide term frequencies: *which specifies how the terms are distributed across the corpus*
  - Inverse Document Frequency (IDF)
- Challenge: a Corpus is dynamic
  - Index, metrics must be updated continuously

Text Classification (I) - "Topic Tagging"

Text Mining

3. Sort the Reviews by Product

Not as straightforward as it seems

"The bPhone-5X has coverage everywhere. It's much less flaky than my old bPhone-4G."

"While I love Acme's bPhone series, I've been quite disappointed by the bEbook. The text is illegible, and it makes even my old Newton look blazingly fast."

"Topic Tagging"

Text Mining

3. Sort the Reviews by Product

Judicious choice of features

- Product mentioned in title?
- Tweet, or review?
- Term frequency
- Canonicalize abbreviations
  - "5X" = "bPhone-5X"

Text Classification (II) Sentiment Analysis

4. Are they good reviews or bad reviews?

- Naïve Bayes is a good first attempt
- But you need tagged training data!
  - The major bottleneck in text classification
  - the main challenge in text classification is getting the tagged data.
- What to do?
  - Hand-tagging
  - Clues from review sites
    - thumbs-up or down, # of stars
  - Cluster documents, then label the clusters

Search and Information Retrieval

Search
&Retrieva
l

5. Marketing team calls up and reads selected reviews in full, for greater insight.

- Marketing calls up documents with *queries*:
  - Collection of search terms
    - "bPhone battery life"
  - Can also be represented as "bag of words"
  - Possibly restricted by other attributes
    - within the last month
    - from this review site

This basically is a search problem, finding the document that meets the search criteria.

Quality of Search Results

Search &Retrieva l

5. Marketing team calls up and reads selected reviews in full, for greate insight.

- Relevance
  - It basically is determining if the results you receive are indeed the ones you want or not.
  - Is this document what I wanted?
  - Used to rank search results
- Precision
  - What % of documents in the result are relevant?
- Recall
  - Of all the relevant documents in the corpus, what % were returned to me?

Computing Relevance (Term Frequency)

Search &Retrieval

5. Marketing team calls up and reads selected reviews in full, for greater insight.

- Assign each term in a document a weight for that term.
- The weight of a term *t* in a document *d* is a function of the number of times *t* appears in *d*.
  - The weight can be simply set to the number of occurrences of *t* in *d* :

$$tf\,(t, d) \;=\; count\,(t, d)$$

  - The term frequency may optionally be normalized.

Inverse Document Frequency (idf)

Search
&Retrieva
l

5. Marketing team calls up and reads selected reviews in full, for greater insight.

$$idf(t) = log\ [N/df(t)]$$

- *N*: Number of documents in the corpus
- *df*(t): Number of documents in the corpus that contain a term *t*

- Measures term uniqueness in corpus
  - "phone" vs. "brick"
- Indicates the importance of the term
  - Search (relevance)
  - Classification (discriminatory power)

TF-IDF and Modified Retrieval Algorithm

5. Marketing calls up and reads selected reviews in full, for greater insight.

- Term frequency – inverse document frequency (tf-idf or tfidf) of term t in document d:

$$tfidf(\mathrm{t, d}) = tf\,(\mathrm{t, d}) * idf(\mathrm{t})$$

query: *brick, phone*

- Document with "brick" a few times more relevant than document with "phone" many times
- Measure of Relevance with tf-idf
- Call up all the documents that have any of the terms from the query, and sum up the tf-idf of each term:

$$\mathrm{Relevance(d)} = \sum_{\mathrm{i} \in [1,\mathrm{n}]} tfidf(\mathrm{t_i, d})$$

Other Relevance Metrics

5. Marketing calls up and reads selected reviews in full, for greater insight.

- "Authoritativeness" of source
  - PageRank is an example of this..rank according to the source
- Recency of document
- How often the document has been retrieved by other users

Effectiveness of Search and Retrieval

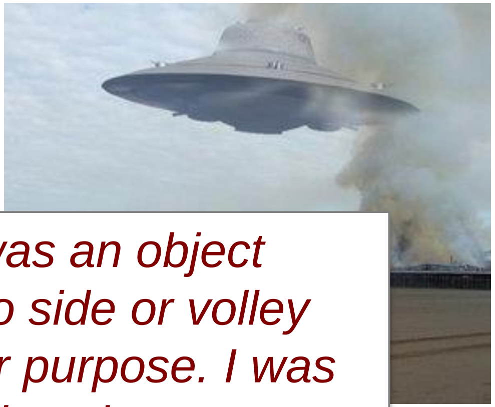
There are other retrieval algorithms

- Relevance metric
  - important for precision, user experience
- Effective crawler, extraction, indexing
  - Crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.
  - important for recall (and precision)
  - more important, often, than retrieval algorithm
- MapReduce
  - Reverse index, corpus term frequencies, idf are implemented effectively with map and reduce algorithms

Natural Language Processing

- Unstructured text mining means extracting "features"
  - Features are structured meta-data representing the document
  - Goal: "vectorize" the documents
- After vectorization, apply advanced machine learning techniques
  - Clustering
  - Classification
    - Decision Trees
    - Naïve Bayesian Classifier
  - Scoring
    - Once models have been built, use them to automatically categorize incoming documents

Example: UFOs Attack



July 15th, 2010. Raytown, Missouri

When I fist noticed it, I wanted to freak out. There it was an object floating in on a direct path, It didn&apos;t move side to side or volley up and down. It moved as if though it had a mission or purpose. I was nervous, and scared, So afraid in fact that I could feel my knees buckling. I guess because I didn&apos;t know what to expect and I wanted to act non aggressive. I though that I was either going to be taken, blasted into nothing, or…

Q: What is the witness describing?

A: An encounter with a UFO.

Q: What is the emotional state of the witness?

A: Frightened, ready to flee.

Source: http://www.infochimps.com/datasets/60000-documented-ufo-sightings-with-text-descriptions-and-metada

Example: UFOs Attack

If we really are on the cusp of a major alien invasion, eyewitness testimony is the key to our survival as a species.

Strangely, the computer finds this account ***unreliable!***

When I fist noticed it, I wanted to freak out
ject
floating in on a direct path, It didn&apos;t move side to side or volley up
n. It moved as if though it had a mission or purpose. I was
nervous, and scared, So afraid in fact that I could feel my knees
buckling. I guess because I didn&apos;t know
d I
wanted to
h that I was either going to be
taken, blasted into nothing, or…

Machine error

Typo

Turn of phrase

Ambiguous meaning

“UFO” keyword missing

Source: http://www.infochimps.com/datasets/60000-documented-ufo-sightings-with-text-descriptions-and-metada

Example: UFOs Attack

Investigators need to…

Search

for keywords and phrases, but your topic may be very complicated or keywords may be misspelled within the document

Manage

document meta-data like time, location and author. Later retrieval may be key to identifying this meta-data early, and the document may be amenable to structure.

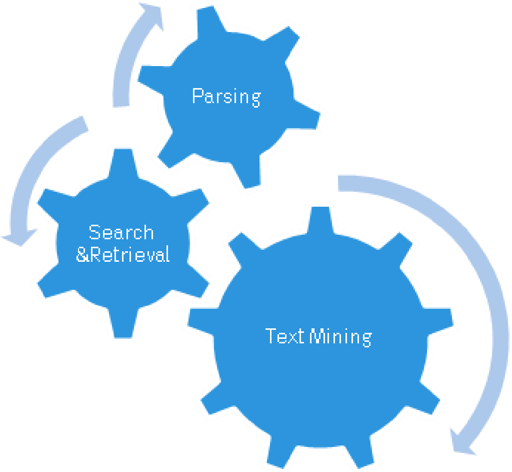
Understand

content via sentiment analysis, custom dictionaries, natural language processing, clustering, classification and good ol' domain expertise.

…with computer-aided text mining

Challenges - Text Analysis

1. Finding the right structure for your unstructured data
2. Very high dimensionality
3. Thinking about your problem the right way



Check Your Knowledge

Your Thoughts?

1. What are the two major challenges in the problem of text analysis?
2. What is a reverse index?
3. Why is the corpus metrics dynamic. Provide an example and a scenario that explains the dynamism of the corpus metrics.
4. How does tf-idf enhance the relevance of a search result?
5. List and discuss a few methods that are deployed in text analysis to reduce the dimensions.

Module 4: Advanced Analytics – Theory and Methods

Part 8: Text Analysis - Summary

During this lesson the following topics were covered:

- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
    - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Use of regular expressions in parsing text
- Metrics used to measure the quality of search results
    - Relevance with tf-idf, precision and recall

Thanks